AD-787 616

SEMANTICS AND SPEECH UNDERSTANDING

Bonnie L. Nash-Webber

Bolt Beranek and Newman, Incorporated

Prepared for:

Advanced Research Projects Agency

October 1974

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R & D

AD 787616

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Bolt Beranek and Newman Inc. 50 Moulton Street Cambridge, Massachusetts 02138	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

Semantics and Speech Understanding

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report

5. AUTHOR(S) (First name, middle initial, last name)

Bonnie L. Nash-Webber

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
October 1974	72	25

8a. CONTRACT OR GRANT NO. DAHC15-71-C-0088 b. PROJECT NO. c. order no. 1697 d.	9a. ORIGINATOR'S REPORT NUMBER(S) BBN Report No. 2896 AI Report No. 19 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) none

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited. It may be released to the Clearinghouse, Department of Commerce for sale to the general public.

11. SUPPLEMENTARY NOTES Reproduced from best available copy.	12. SPONSORING MILITARY ACTIVITY Advanced Research Projects Agency 1400 Wilson Boulevard Arlington, Virginia 22209

13. ABSTRACT

In recent years, there has been a great increase in research into automatic speech understanding, the purpose of which is to get a computer to understand the spoken language. In most of this recent activity, it has been assumed that one needs to provide the computer with a knowledge of the language (its syntax and semantics) and the way it is used (pragmatics). It will then be able to make use of the constraints and expectations which this knowledge provides, to make sense of the inherently vague, sloppy and imprecise acoustic signal that is human speech.

Syntactic constraints and expectations are based on the patterns formed by a given set of linguistic objects, e.g. nouns, verbs, adjectives, etc. Pragmatic ones arise from notions of conversational structure and the types of linguistic behavior appropriate to a given situation. The bases for semantic constraints and expectations are an a priori sense of what can be meaningful and the ways in which meaningful concepts can be realized in actual language.

We will attempt to explore two major areas in this paper. First we will discuss which of those things that have been labelled "semantics", seem necessary to understanding speech. From the opposite point of view, we will then argue for speech as a good context in which to study understanding. To illustrate these points, we will begin by describing, albeit briefly, how semantics is being used in several recent speech understanding systems. We will then expand the generalities of the first section with a detailed discussion of some actual problems that have arisen in our attempt to understand speech.

DD FORM 1 NOV 65 1473 REPLACES DD FORM 1473, 1 JAN 64, WHICH IS OBSOLETE FOR ARMY USE.

Unclassified

Security Classification

Unclassified

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Artificial Intelligence						
Automatic Speech Understanding						
Case Frames						
Computational Linguistics						
Computational Semantics						
Natural Language Processing						
Semantic Networks						
Semantics						
SPEECHLIS						
Speech Understanding						
Speech Understanding Research						
Speech Understanding Systems						

1a

Unclassified

Security Classification

BBN Report No. 2896
AI Report No. 19

Bolt Beranek and Newman Inc.

SEMANTICS AND SPEECH UNDERSTANDING

BONNIE L. NASH-WEBBER

OCTOBER 1974

This research was supported by the Advanced Research Projects Agency of the Department of Defense under ARPA Contract No. DAHC15-71-C-0088 and ARPA Order No. 1697.

ib

TABLE OF CONTENTS

Abstract

In recent years, there has been a great increase in research into automatic speech understanding, the purpose of which is to get a computer to understand the spoken language. In most of this recent activity, it has been assumed that one needs to provide the computer with a knowledge of the language (its syntax and semantics) and the way it is used (pragmatics). It will then be able to make use of the constraints and expectations which this knowledge provides, to make sense of the inherently vague, sloppy and imprecise acoustic signal that is human speech.

Syntactic constraints and expectations are based on the patterns formed by a given set of linguistic objects, e.g. nouns, verbs, adjectives, etc. Pragmatic ones arise from notions of conversational structure and the types of linguistic behavior appropriate to a given situation. The bases for semantic constraints and expectations are an a priori sense of what can be meaningful and the ways in which meaningful concepts can be realized in actual language.

We will attempt to explore two major areas in this paper. First we will discuss which of those things that have been labelled "semantics", seem necessary to understanding speech. From the opposite point of view, we will then argue for speech as a good context in which to study understanding. To illustrate these points, we will begin by describing, albeit briefly, how semantics is being used in several recent speech understanding systems. We will then expand the generalities of the first section with a detailed discussion of some actual problems that have arisen in our attempt to understand speech.

I. INTRODUCTION

Psychologists have demonstrated that it is necessary for people to be able to draw upon higher level linguistic and world knowledge in their understanding of speech: the acoustic signal they hear is so imprecise and ambiguous that even a knowledge of the vocabulary is insufficient to insure correct understanding. For example, Pollack and Pickett's experiments [11] with fragments of speech excised from eight-word sentences and played to an audience showed that 90% intelligibility was not achieved until a fragment spanned six of the eight words, and its syntactic and semantic structures were becoming apparent. (See Wanner [19] for an excellent survey of psycholinguistic experiments on listening and comprehension.) Similarly, the apparent impossibility of building a "phonetic typewriter" [10] (a machine which types on paper the words and sentences spoken into it) or of extending systems capable of single-word recognition to ones capable of recognizing continuous speech seems to imply that this ability to draw on higher level knowledge is necessary for computers too.

That a person's expectations, based on his linguistic and world knowledge, often have a stronger influence on what he hears than the actual acoustic signal is also evidence for the strong part that this knowledge plays in understanding speech. Anecdotes illustrating this point

abound. For example, at the recent IEEE Symposium on Speech Recognition, the projectionist repeatedly heard the speaker request "House lights, please", when what had been said was "Next slide, please". The projectionist seemed to expect that a speaker would not keep the auditorium dark for so long. When the acoustic signal is ambiguous out of context, as the signal equally understandable as "his wheat germ and honey" or "his sweet German honey", it is only higher level knowledge which can force one reading over the other.

Without making any claims about how a person actually understands speech, this paper will discuss some aspects of semantic knowledge and how they seem to contribute to understanding speech. From the opposite point of view, we will argue why speech is a good context for studying understanding. For illustration, we will present a brief description of how semantics is being used in several recent speech understanding systems, and conclude with a more detailed description of the part semantic knowledge plays in SPEECHLIS, a speech understanding system being developed at Bolt Beranek and Newman.

II. ASPECTS OF SEMANTIC KNOWLEDGE FOR AUTOMATIC SPEECH UNDERSTANDING

If a speech understander must use semantic knowledge to constrain the many possible ways of hearing an utterance, then his semantic knowledge must represent what can be meaningful and what may be expected at any point in a dialogue. Preferring a meaningful and likely utterance to one that is not, a speech understander must be able to use his semantic knowledge to seek one out. Thus it is the knowledge of what can be meaningful and the ability to make predictions based on that knowledge that may be the most important aspects of semantics for speech understanding. As to the former, it is more important to know that physical objects can have color than that canaries are yellow. As to the latter, if the objects in a group can be distinguished by color, then it is reasonable to expect a color specification in identifying a subset of them. This makes "yellow birds", for example, a meaningful and likely phrase. This is not to say that factual knowledge is not useful in speech understanding, but rather, as we hope to show below, that it is just not as powerful an aid as other types of semantic knowledge. Let us now consider what types of semantic knowledge determine what is meaningful and enable predictio ?

A. Knowledge of Names and Name Formation

Semantic knowledge of the names of familiar things and of models for forming new ones permits a listener to expect and hear meaningful phrases. For example, knowing the words "iron" and "oxide" and what they denote, and that a particular oxide (or set of them) may be specified by modifying the word "oxide" with the name of a metal may enable a listener to hear the sequence "iron oxides", rather than "iron ox hides" or even "Ira knocks sides".

B. Knowledge of Lexical Semantics

Knowledge of lexical semantics (models of how words can be used and the correspondence between concepts in memory and their surface realizations) enables the listener to predict and verify the possible surface contexts of particular words. Along with the previously mentioned knowledge of names and name formation, this contributes to "local" recognition of an utterance: given a hypothesis that a word has occurred in the utterance, what words could have appeared to its left or right. For example, the concept of CONTAINMENT, invoked, inter alia, when the word "contain" appears in a sentence, has two other concepts strongly associated with it - a container and a containee. (These might also be called the "arguments" to CONTAINMENT. Note, in this paper, concepts will be distinguished from words by

being written in capital letters.) When "contain" is used in an active sentence, it must have a subject which is understood to be a location or container, and an object which is capable of being located or contained. In a passive sentence, the roles are interchanged: the active object becomes the passive subject and the active subject or location is realized in a prepositional phrase headed by "in".

Every egg contains a yolk.
(Active)
A yolk is contained in every egg.
(Passive)

There are several things to notice here. First, given the possibility of being able to hear the initial segment of the first utterance as either "every egg" or "every ache", one would usually hear the former, since it is a more likely container, especially for yolks. Secondly, given that little words lose most of their phonetic identity in continuous speech and that in hearing the second utterance we have a strong hypothesis that it is of a passive sentence, we can use the knowledge of how "contain" passivizes to predict and verify the occurrence of "is" and "in" in the acoustic signal. If we cannot satisfy ourselves as to their existence in the utterance, we may decide to change our earlier hypothesis that the utterance was of a passive sentence.

Thirdly, while we can profitably use lexical semantics to predict the local context of a word by going to the concepts it can partially instantiate and predicting what can fill the gaps, it does not gain one much to make predictions about the way in which a completely uninstantiated concept will be realized. There are usually too many possibilities available. For example, the concept of CONTAINMENT comes across in all the following phrases:

Rocks containing sodium
Sodium-containing samples
Sodium-rich basalts
Igneous samples with sodium
Samples in which there is sodium
Rocks which have sodium

C. Knowledge of Conceptual Semantics

Knowledge of conceptual semantics, how concepts are associated in memory, contributes to a listener's ability to make "global" predictions across utterances, as well as ones local to a given one. The global predictions are primarily of the nature: if one concept is under discussion, which other ones are soon likely to come up and which ones not. Expectations about which related concepts need not be mentioned in the discourse help the listener accept and accommodate such discourse tricks as ellipsis and anaphora. A short example of conversation should suffice here to illustrate the point.

"I'm flying to New York tomorrow.
Do you know the fare?"
"About 26 dollars each way."
"Do I have to make reservations?"
"No."
"Super."

There are several points to make here. First, the concept of a trip is strongly linked with such other concepts as destinations, fares, transportation mode, departure date, etc. So one might expect them to be mentioned in the course of a conversation about a trip. Secondly, the strength of these associations is both domain-, context- and user-dependent. If the domain concerns planning trips, as in making airline reservations, then destination and departure date would seem to have the strongest links with trips. In another domain such as managing the travel budget for a company, it may only be the cost of the trip and who is paying for it that have this strong association. As far as context and user dependency, the company accountant's primary interest in business trips may be quite different from that of a project leader wondering which of his people is going where.

Thirdly, the places where ellipsis is most likely to occur seem to correlate well with strong inter-concept associations. This is useful information since it suggests when not to look hard for related concepts in the local context. For example, "the fare" and "reservations" are both elliptical phrases: "the fare" must be for some trip

via some vehicle at some time. ut fares are so strongly linked with these notions that is is not necessary to mention them explicitly as in, "Do you know the current air fare to New York?" Again, what the reservations are for is not stated explicitly, but must also be for the aforementioned flight. Without a knowledge of the concepts associated with trips and fares and how "strong" the links are, none of the above local or global predictions could be made. What's more, the above conversation would be incoherent. (N.B. Conceptual associations such as those discussed above are of course not the only source of "global expectations". Rhetorical devises available to a speaker who chooses to use them, such as parallelism and contrast, add to global expectations about the structure of future utterances. In addition, problem solving situations also have a strong influence on the nature of discourse and the speaker's overall linguistic behavior.)

D. Knowledge of the Use of Syntactic Structures

Knowledge of the meaningful relations and concepts that different syntactic structures can convey enables the listener to rescue cues to syntactic structure which might otherwise be lost. Among the meaningful relations between two concepts, A and B, that can be communicated syntactically are that B is the location of A, the possessor of A, the agent of A, etc. Also among syntactically

communicated concepts are set restriction (via relative clauses), eventhood (via gerund constructions), facthood (via 'that'-complements), etc. Syntactic structure is often indicated by small function words (e.g. prepositions and determiners) which have very imprecise acoustic realizations. The knowledge of what semantic relations can meaningfully hold between two concepts in an utterance and how these relations can be realized syntactically can often help in recovering weak syntactic cues. On the other hand, one's failure to recover some hypothesized cue, once attempted, may throw doubt on one's semantic hypothesis about the utterance. For example, the preposition "of" can practically disappear in an utterance of "analyses of ferrobasalts". Yet the only meaningful relation between "analyses" and "ferrobasalts" that can be expressed with this word order requires that "ferrobasalts" be realized as a prepositional phrase headed by "of" or "for". If one hypothesizes that something is an utterance of "analyses of ferrobasalts", and one is reasonably certain only that he has heard "analyses" and "ferrobasalts", he can try to confirm the occurrence of one of these prepositions in the speech signal. If he can, it is more believable that "analyses of ferrobasalts" was the intended sentence. If he can not, it becomes doubtful, though not impossible, that "analyses" and "ferrobasalts" really did occur in the utterance. An alternative hypothesis, for example, that the

intended sentence was "analyses for all basalts", may become more likely.

E. Knowledge of Specific Facts and Events

Knowledge of specific facts and events can also be brought in as an aid to speech understanding, though it is less reliable than the other types of semantic knowledge discussed above. This is because it is more likely for two people to share the same sense of what is meaningful than for them to be aware of the same facts and events. Fact and event knowledge can be of value in confirming, though not in rejecting, one's hypotheses about an utterance. For example, if one knows about Dick's recent trip to Rhode Island for the Americas Cup, and one hears an utterance concerning some visit Dick had made to -- Newport?, New Paltz?, Norfolk?, Newark? -- one would probably hear, or chose to hear, the first, knowing that Dick had indeed been to Newport. However, one couldn't reject any of the others, on the grounds that the speaker may have more information than the listener.

III. WHY STUDY SEMANTICS IN THE CONTEXT OF SPEECH?

In our attempt to do automatic speech understanding, we have become aware of aspects of the language understanding process that either haven't arisen in the attempt to understand printed text, or have done so and been consciously put aside as not crucial to the level of understanding being attempted.

The first aspect concerns the nature of the input. In spoken language, as distinct from written text, word boundaries are not given unambiguously, and hence words may not be uniquely identified. Compounding the problem is the sloppy, often incomplete realization of each word. In addition, coarticulation phenomena are such that the correct identification of a word in the speech signal may depend on the correct identification of its neighbors. Conversely, a word's incorrect identification may confound that of its neighbors.

As a result of the nature of its input, understanding spoken language seems to require a special mode of operation, such as "hypothesize and test", in order to get around the vague, often incomplete, realization of each word in the utterance. That is, one needs the ability to make hypotheses about the content of some portion of the input and then verify that that hypothesis is consistent with a

complete interpretation of the input. The same process must go on in the understanding of handwritten text, which is inevitably sloppy and ill-formed. Notice, for example, how the same scrawl is recognized as two different words in contexts engendering different predictions.

Pole vaulting was the third event of the meet.

After dinner, Jack went home.

Recently, researchers concerned with modelling human language understanding, notably Riesbeck [13], have also proposed this mode of operation, "parsing with expectations", as the way of getting directly to, in most cases, the "intended" interpretation of a sentence. His argument is that this model accounts for the fact that people do not even seem to notice sense ambiguities if they are expecting one particular sense.

The second aspect of language understanding that comes out through the context of speech is that there appear to be better and worse readings of a sentence or utterance, as opposed to good and bad ones. In speech understanding, we are no longer up against the problem of rejecting bad readings of

"I saw the Grand Canyon flying to New York."

(E.g. the one in which the Grand Canyon is doing the flying), but rather in choosing among such possible readings of an imprecise acoustic signa as:

> How many people like ice cream?
> Do many people like ice cream?
> Do any people like ice cream?
> Do eighty people like ice cream?

Some are "better" than others: one is forced into weighing many factors in choosing the best -- closeness of some realization of the reading to the acoustic signal, appropriateness of the reading to the context, likelihood of the reading within the context, etc. And all the factors may not point to the same reading as being best.

The next point about the advantages of studying understanding in the speech context is that there are phenomena relevant to understanding which are found either exclusively in spoken language, or mainly there and only rarely in written text.

One of these is the marvelous errors in speech production which, though funny, must still be accounted for in any valid model of human language understanding. The errors occur at all linguistic levels -- phonemic, syntactic, semantic -- and include such effects as spoonerisms, malapropisms, portmanteaus, mixed metaphors and idioms, etc. For example,

> "I'm glad you reminded me: it usually takes a week for something to sink to the top of my stack."

> "Follow your hypothesis to its logical confusion."

(See Fromkin [6] for additional examples.) These errors rarely occur in text, whose production is much more deliberate and considered than that of speech. They should be rigorously studied, since they force a constraint on valid models of human language organization which correct linguistic behaviour does not.

Another of these phenomena is that of stress, intonation, and phrasing. Though many linguists would argue that they are regularly predictable on the basis of the syntactic structure of the utterance alone, I would agree with Bolinger [3] that these are not only syntactic phenomena, but are also used by a speaker to reflect his intended meaning and focus. Thus, to quote two of Bolinger's examples, the difference in stress patterns between the two utterances shown below cannot be accounted for on the basis of syntactic structure, which is the same for both, but reflects a difference in information focus.

> The end of the chapter is reserved for problems to computerize.

> The end of the chapter is reserved for problems to solve.

"Computerize" is richer in meaning than simply "solve". The

choice of the former verb, rather than the latter, seems to reflect a decision that the action, not the object (i.e. "problems") is the point of information focus. The difference in intonation reflects this choice.

There are two points here: first, it is possible in speech to have several different, but simultaneous cues to the same information. For example, potential ambiguities in the scopes of prepositional phrases may never arise because of semantic constraints or contextual knowledge or appropriate intonation or phrasing. It is an interesting question whether or not a speaker actually uses all possible cues if fewer will suffice to resolve a potential ambiguity. More generally, there are factors which any model of human language understanding must account for, like the ones above, which can only be studied in the context of speech.

Finally, the attempt to understand speech forces us to confront and deal with what we consider one of the most important and difficult to understand aspects of any decision process, and that is the role of error analysis and correction. We mentioned earlier the inherently ambiguous nature of the input. Given we have decided that our reading of part or all of an utterance must be wrong, we must be able to suggest where the source of the error lies and what the best alternative hypothesis is. Moreover we must do so efficiently, lest we fail to come up with a satisfactory

reading in reasonable time. These problems of error analysis and correction have been the focus of a great deal of past, present and future research in Artificial Intelligence, research which is being avidly followed by the speech understanding community. (See references 6,12, 15, 19 and 21 for several different schemes for dealing with these problems.)

IV. THE STATE OF SEMANTICS IN AUTOMATIC SPEECH UNDERSTANDING

Rollin Weeks, of the SDC Speech Understanding Project [1,14], once called automatic speech understanding the coming together of two unsolved problems -- speech and understanding. It is no wonder then, that in the short time that this field has been actively researched [9], very little has been done on semantics and automatic speech understanding per se. There is just too much to be done on the many other aspects of the field, aspects like phonetic segmentation and labelling of the speech wave which promise an earlier reward. (At the time of this writing, all the systems which will be discussed here - CMU's HEARSAY, SDC's VDMS, Lincoln's VASSAL, that of SRI, and BBN's SPEECHLIS - were undergoing major changes. While these new systems promise to be very interesting, exciting and even successful, this survey will only cover the earlier versions of those systems which we can reference.)

A. The HEARSAY I System

In the HEARSAY I speech understanding system developed at Carnegie-Mellon University [9,12], semantics refers to both general knowledge about the system's task and specific knowledge about its current state. In this case, the task involves a chess game in which HEARSAY is one of the players

and the human speaker, the other. The dialogue consists of each player speaking aloud his moves. Each time a move spoken by the user is to be recognized, the semantics module receives a list of legal moves, ordered by goodness, from the chess playing component of the system. This list of legal moves, which reflects both general knowledge about chess and specific knowledge about the state of the board and the best moves under the circumstances, is used for several purposes: to make hypotheses about what HEARSAY's opponent, the speaker, has said; to make specific proposals as to what words are likely to have occurred where in the utterance; and to pare down the set of words proposed by other modules to ones consistent with legal moves.

The language spoken and recognized by HEARSAY I is very limited. Consequently its syntax and semantics can be highly constrained and still be completely adequate to the task. The language basically involves a one-for-one reading of standard chess notation in the order in which the symbols occur, with some ellipsis being permitted in specifying the piece being moved or the square being moved to. Thus, the move QRP/QR2-QR4 can be read, with or without modifiers, as most anything from the fully expanded form:

Queen (or Queen's) Rook (or Rook's) <u>Pawn</u> on Queen (or Queen's) Rook 2 <u>Goes-to</u> (or Moves-to or To) Queen (or Queen's) <u>Rook</u> 4

to the minimal form:

<u>Pawn To Rook 4</u>

There are finitely many sentences in the language, and the list of legal moves and the rules for ellipsis constitute the only semantic data that the system has. Working from left to right then, based on words already recognized in the utterance, semantics notes which legal moves these words are consistent with. For each one, it makes predictions of what might follow the rightmost word, taking into account the possibility of ellipsis and the goodness of move. When syntax makes a set of similar predictions based on the allowable sentences of the language, semantics can use its knowledge of currently meaningful sentences (legal moves) to pare down the set. Note that it is not similarly advantageous for syntax to screen semantics' predictions because, given the way semantics is organized, every semantically meaningful sentence is syntactically correct. Moreover, it is impossible for HEARSAY I to hear an illegal move (or even an implausible one).

What aspects of semantic knowledge useful for speech understanding does HEARSAY I semantics display? Basically, it has a knowledge of everything that can be said meaningfully, and, with its rules of ellipsis and its list

of legal moves, a knowledge of all possible ways to say each thing. It uses these aspects of semantics actively, both to predict and to constrain the possibilities. Thus in its limited domain, it takes advantage of everything semantics has to offer.

B. A Vocal Data Management System

In the vocal data management system (VDMS) developed at System Development Corporation [1,14], interest also rests in drawing out as much as possible from a highly constrained syntax and semantics. A VDMS user interacts with the system in an artificial, but English-like data management language to access information on the submarine fleets of the United States, Soviet Union and United Kingdom. Typical requests to VDMS would include:

> Total quantity where type equals nuclear and country equals USA.
>
> Print type where missiles greater than seven.

The data management language can be described by about 35 "recursive syntactic equations" (production rules) which reference both syntactic/semantic categories like "item name" (e.g. "country", "type") and "item value" (e.g. "USA", "nuclear"), and syntactic terminals (e.g. "print", "total"). Thus, such sentence-level semantics as knowledge

of the sentential context in which a given word can occur is inherent in the grammar and is handled by the VDMS parser. This knowledge is used both top-down and bottom-up to determine the grammatical structure of the utterance and to predict both words and phrases adjacent to a given one.

The other chunk of semantic knowledge in VDMS, comprising knowledge of what things are being discussed, how they can be expressed linguistically, and what tasks can be pursued, is located in the "discourse level controller". Along with the parser, these two form the VDMS linguistic controller. The discourse level controller is itself divided into two modules: a user model and a thematic memory. The former determines the "state" the user is in (e.g. interactive query mode, user aid mode), and from it, predicts likely "syntactic equations" for the next user interaction. For instance, in interactive query mode, those "equations" for Print, Repeat, Count, and Total are all likely. Thematic memory is concerned with the content of the utterance rather than its type: in particular, it anticipates the content words that might occur somewhere in the next utterance. It does this by maintaining a dialogue history: for each word that occurs in either question or answer, it assigns a weight based on the number of times it has occurred in the last several interactions and the manner of its use (e.g. as a value in an answer, as an unqualified name in a question, etc.). Words with high weights are

proposed as highly likely to occur somewhere in the next utterance and are actively sought. There does not appear to be any attempt to predict how they might be used though, or to predict logically related words or concepts. Moreover it is not clear from their experiments whether this particular use of thematic memory helps or hinders their performance.

Semantics and syntax -- since it is difficult to separate the two in the SDC system -- are being exercised in a very strong predictive capacity. This is acknowledged in their calling their strategy of employing the domain-dependent syntactic and semantic constraints PLC, "Predictive Linguistic Constraints". While it is not clear that the specific strategies being employed in the user model and thematic memory modules are of any theoretical interest, the desire and the ability to drive the recognition process both top-down -- predicting the general form and content of the next utterance -- and bottom-up -- predicting adjacent words from ones already recognized -- do take best advantage of those things that semantics and the other higher level knowledge sources have to offer.

C. VASSAL

Like the two systems discussed previously, the Lincoln speech understanding system [5] is attempting to understand speech in a limited domain. In this case, it is to enable

vocal control of a system for studying the acoustic correlates of phonemic events. A researcher can use the system to search the Lincoln speech data base, take measurements on its acoustic data, and display the data, any measurements on it, and the results of any tests done on those measurements. The subset of English in which he can converse with the system was compiled from a study of the subtasks he might want to accomplish and the ways he would request their execution. Such subtasks include specifying the phonemic sample, specifying the measurements to be taken, running a statistical discrimination program, examining the results, and using the display programs. Each utterance to the system is a command containing information relevant to only one subtask at a time. In most cases, the utterance also contains <u>all</u> the information relevant to that subtask, so that a dialogue with the system consists of a series of subtask requests, each followed by the system's execution of it. These subtasks are only loosely organized within the single main task of studying the acoustic correlates of phonemic events. The result is that the system can make no predictions about future utterances based on past ones. Sample requests to the system include:

> Display the formant graph on the Hughes scope.
>
> Erase the display of the confusion matrix.
>
> Recompute the average energy in the second voiced segment.

The Lincoln system consists basically of three modules: a phonetic recognition module acting both as an acoustic front end and as a phonetic hypothesis verifier, a linguistic module, and a functional response module. (VASSAL is only one of three linguistic modules capable of being hooked into the system, but our discussion here will be confined to its VASSAL configuration as VASSAL has been used and reported on most often as the higher-level component of their system.) Again, as in the previous systems, VASSAL uses a finite categorial grammar (i.e. one whose non-terminal elements are semantic, rather than syntactic, classes) to make hypotheses about the content of the utterance. Input to VASSAL is a sequence of acoustic-phonetic elements (APELs) produced by the phonetic recognition module in its front end capacity, from an analysis of the speech waveform. After receiving this input, VASSAL works in a top-down and left-to-right manner, making hypotheses about the next word to be recognized from the semantic classes permitted there by the grammar. These hypotheses are then judged for adequacy of fit against the acoustic signal by the phonetic recognition module in its verification mode. The result of this processing is VASSAL's best reconstruction of the utterance and an interpretation in terms of what functional response for what arguments has been requested.

Note that there is an important difference here between CMU's system, HEARSAY, and Lincoln's. While it was impossible for HEARSAY to hear a sentence inconsistent with the state of its world model (i.e. a move not allowed in the chess game), it is possible for such a situation to arise in the Lincoln system. For example, VASSAL may decide its best hypothesis about the original utterance was "Skip to the thirtieth sentence on tape unit six". If there were only twenty sentences stored on that tape unit, the above reconstruction would be linguistically acceptable, but nevertheless inconsistent. Deciding what to do about this inconsistency was planned as one of the responsibilities of the functional response module. Either the user actually said something wrong or there was an error in the linguistic module's reconstruction. The functional response module was to make available to the linguistic module information about the inconsistency (a non-trivial matter), so the linguistic module could decide whether its first choice was wrong and a second best one would not result in the inconsistency, or else whether the user has asked something impossible. However, work on the Lincoln speech understanding system was terminated before much was done on this problem.

D. A Natural Language Speech Understander

In the speech understanding projects discussed so far, the objects of interest have only been artificial

English-like languages, which are nevertheless sufficient for their respective task domains. The group at the Stanford Research Institute however, rather than constraining the language a priori, is attempting to deal with whatever linguistic constructions are natural within their vocabulary and task domain. This has been true of all the versions of their speech understanding system, though only their first will be discussed here. This system [17,18] had its higher level organization and knowledge based on Winograd's "Computer Program for Understanding Natural Language" [21] and was designed for such utterances about SHRDLU and his toy blocks world as:

> Put the black block in the box.
> How many green pyramids are on top of blocks.

Basically, the system was driven, top-down, left-to-right, by the parser. The parser could, in principle, call upon all sorts of syntactic, semantic, pragmatic and inferential knowledge to guide it through the grammar to a complete recognition of the utterance, this by predicting and constraining the possible words at each successive word boundary. When the higher-level knowledge sources came up with a set of words possible at a given word boundary, a word verification program would match each word against the appropriate portion of the utterance and return, for each, a goodness of match score. If some word matched well enough, the system would continue to follow its current

path through the grammar, adding the new words to the string of already recognized ones. Otherwise, it would backtrack and try again, until it came up with an acceptable reconstruction of the utterance.

Examples of the kinds of semantic knowledge employed, or planned for, included:

a) lexical semantic (case frame) information about verbs: the number of arguments (cases) that a verb takes, what types of things can fill each case, and how the verb and its arguments can appear in an utterance. This can be used to order paths through the grammar or constrain the words proposed.

b) semantic marker information -- used with case frame information to predict and verify what can fill a particular case. This is also used to discover which kinds of nouns can be modified by which already recognized adjectives and nouns.

c) information about the natural ordering of modifiers of a noun, from most to least adjective-like. For example, "the big red ball" is a more likely noun phrase than "the red big ball".

Unfortunately, this first version of SRI's speech understander was never fully implemented, so we cannot evaluate its success. Yet it would seem that, while the system was organized to take advantage of much that semantics and other higher-level knowledge sources have to offer and use it for both hypothesis generation and verification, much more was lost by not relying more on the acoustic signal to formulate some initial hypotheses. As we mentioned in the section on lexical semantics, there are usually too many possibilities when hypothesizing is virtually unconstrained. The version of their system now

being developed will work both top-down and bottom-up in the process of generating and verifying its hypotheses.

E. SPEECHLIS

Like the group at SRI, the SPEECHLIS project at Bolt Beranek and Newman [23,15] has taken on the more ambitious task of attempting to understand the kind of English natural to a particular domain and vocabulary. Currently, there are two domains in which the problems of speech understanding are being studied: a natural language information retrieval system for lunar geology and a similar one for travel budget management. Typical requests in the two domains include:

<u>Lunar Geology</u>
- Give the average K/Rb ratio for each of the fine-grained rocks.
- Which samples contain more than 10ppm aluminum?
- Has lanthanum been found in the lunar fines?

<u>Travel Budget Management</u>
- How much was spent on trips to California in 1973?
- Who's going to the ACL meeting in June?
- If we only send two people to IFIP-74, will we be within our budget?

It is envisioned that a user will carry on a spoken dialogue with the system in one of these areas in order to solve some problem.

The reason for choosing the former area was to draw upon a two-year experience with the BBN LUNAR system [23]. The latter permits investigation of the problems of user and

task modelling, which turn out to be very inconvenient in the specialized technical area of lunar geology. The lunar geology vocabulary for the SPEECH system contains about 250 words, of which approximately 75 are "function words" (determiners, prepositions, auxiliaries and conjunctions) and the remaining 175 are semantically meaningful, "content words". The vocabulary for travel budget management is larger, containing about 350 words, with approximately the same core of 75 function words.

In SPEECHLIS, semantics refers in the generative direction to a knowledge of concepts, the meaningful associations between them, and their possible lexical and syntactical realizations in an utterance. In the analytic direction, semantics also refers to a knowledge of what concepts are completely or partially instantiated by any given word and what relationships may be expressed through the use of any given syntactic construct (Although the representation of this latter knowledge has not yet been completed). It also refers to such domain specific facts, as for example (in the travel budget management system): John delivered a paper at the ICA conference in London in July, 1974. All this knowledge is represented in a semantic network and in frames attached to nodes in the network.

In the recognition strategy documented in [15] and [23], this semantic knowledge is used in several ways. For

each word match (in a strategy which is not left-to-right, but rather involves looking for all words which match some region of the utterance very well), semantic knowledge is used to make predictions about the local context of the word based on the concepts fully or partially instantiated by it. Semantic knowledge is then also used to form sets of good word matches, all of which fit together meaningfully to fully or partially instantiate some further concept. These sets, along with the semantic motivations for forming them, are semantics' hypotheses about the original utterance. When it is profitable to check whether some possible syntactic organization of the set of word matches would substantiate semantics' hypotheses, the set is sent to syntax for any structuring it can propose. Semantic knowledge of the possible syntactic realizations of concepts is used to do this checking. The same semantic knowledge may also be used by syntax for efficiency reasons in guiding its hypotheses. If there are several ways of syntactically structuring the set of word matches, the ones which are consistent with the semantic hypothesis should be the ones proposed first. Together, aided by a knowledge of utterances likely in an information retrieval environment, semantic and syntactic knowledge are used to reconstruct the input utterance. In the version of SPEECHLIS documented in references [15] and [23], no attempt had yet been made to make cross-utterance predictions based on a user or context

model.

F. Summary

Looking at these five speech understanding systems, one sees more similarities than dissimilarities in their view of syntactic, semantic, and world knowledge. First, they all aspire to make active use of this knowledge, using its constraints on acceptable and appropriate utterances to work outwards from what they are relatively sure of to what is initially more vague and ambiguous. That some of the systems work from left-to-right in this process while others work from whatever strong "anchor points" they can identify, is more a difference of implementation than of philosophy. Secondly, most of the systems aspire to making predictions about the likely content of an incoming utterance before analyzing the utterance itself. This intuitively seems to reflect human speech understanding. Unfortunately, none of them yet seems to do it very well. Thirdly, all of the systems have chosen one or two small task domains in which to consider their spoken input. What's more, all these tasks are interactive: the user speaks and the system is counted upon to respond. This is quite different from the early, naive vision of a speech recognition system passively recording everything said to it, a vision embodied in the phonetic typewriter mentioned earlier.

V. SPECIFIC SEMANTIC PROBLEMS IN SPEECH UNDERSTANDING

We shall now attempt to give a more detailed discussion of how a speech understander might use a knowledge of meaningful concepts and their possible surface realizations in order to recover a speaker's intended utterance. This discussion will be in terms of SPEECHLIS, the BBN Speech Understanding System, because of its large aspirations and our own familiarity with it. Its value will be in pointing out many interesting specific problems in speech understanding, in more concrete terms than the generalities the first half. The deficiencies we note in our current solutions may help the reader avoid having to discover these deficiencies for himself, and may suggest to him better solutions. In addition, the framework presented may be suggestive to psychologists and psycholinguists attempting to discover and explain the mechanisms by which humans understand language.

Before discussing semantics and speech understanding in terms of SPEECHLIS then, it will be useful to describe in more detail the structure of SPEECHLIS and, hence, the kinds of information available for making and verifying semantic hypotheses.

A. The SPEECHLIS Environment

Before formulating our design for SPEECHLIS, we attempted to get an understanding and intuitive feel for the nature of the speech understanding task -- the different sources of knowledge that would be necessary; the situations in which different kinds of knowledge should be brought to bear, the kinds of inferences required; and the interactions among the different knowledge sources. We did this by means of a series of "incremental simulations" of a complete speech understanding system. With this technique, well described in Woods and Makhoul [25], a person simulates a part of a system which is not yet formulated in order to gain insight into how it might work. At the time of their paper [24], the only part of SPEECHLIS already machine-implemented was an embryonic word matcher which could match phonetic word spellings against a region of a partial phonetic transcription of an utterance and a lexical retrieval package which could look at a user specified region of the utterance and return a list of words for which at least one possible phonetic spelling might match in that region. This list would suggest to the user what words to send as input to the word matcher. (N.B. The distinction we are making is that word matching is a top-down predictive process, whereas lexical retrieval is a bottom-up data-driven one.) Such sources of knowledge as syntax, semantics and pragmatics were being simulated by a person.

The initial design of SPEECHLIS was largely based on our sense of what was going on during these incremental simulations and on the problems which presented themselves there. [For a more detailed exposition of the SPEECHLIS world, see references 15 and 23.]

In order to establish the environment in which the semantics component of SPEECHLIS operates, we will now give a brief description of the SPEECHLIS world as it evolved through our incremental simulations.

Because the inherent ambiguity and vagueness of the speech waveform do not permit a unique characterization of it by acoustic-phonetic programs, it is possible for many words to match, to some degree, any region of the input. The solution we came up with to this problem was to have the lexical retrieval and word matching programs produce a <u>word lattice</u>, whose entries were words which were found to match well (i.e. above some threshold) in some region of the utterance. Associated with each such word match was a description of how and how well it matched the input. There was also the poblem that small words like "an", "and", "in", "on", etc. tend to lose their phonemic identity in speech and result in spurious matches everywhere. This we avoided by initially trying to match only words of three of more phonemes in length. The motivation for looking at all strong matches at once, rather than accessing them in a

strict left-to-right way, was basically efficiency. In our incremental simulations, we found there were just too many syntactic and semantic possibilities if we couldn't use the really good matches of content words to suggest what the utterance might be about.

This initial, usually large lattice of good big word matches then serves as input to the syntactic, semantic, and pragmatic components of the system. Subsequent processing involves these components working, step by step, both separately and together, to produce a meaningful and contextually apt reconstruction of the utterance, which is hoped to be equivalent to the original one. We noticed in our incremental simulations that most often our actions in proposing or choosing a word reflected some hypothesis about what the original utterance might be. In SPEECHLIS, this notion of a current hypothesis appears as the object we call a <u>theory</u>. Each step in the higher-level processing of the input then is highlighted by the creation, evaluation, or modification of a theory, which is specifically a hypothesis that some set of word matches from the word lattice is a partial (or complete) reconstruction of the utterance.

The word lattice is not confined, however, to the initial set of "good, long" word matches. During the course of processing, any one of the higher level components may make a <u>proposal</u>, asking that a particular word or set of

words be matched against some region of the input, usually adjacent to some word match hypothesized to have been in the utterance. The minimum acceptable match quality in this case would be less than in the undirected matching above for two reasons. First, there would be independent justification from the syntax, semantics, and/or pragmatics components for the word to be there, and second, the word may have been pronounced carelessly because that independent justification for its existence was so strong. For example, take a phrase like "light bulb", in ordinary household conversation. The word "light" is so strongly predicted by bulb in this environment, that its pronunciation may be reduced to a mere blip that something preceded "bulb". In the case of proposals made adjacent to, and because of, some specific word match, the additional information provided by the phonetic context of the other word match will usually result in a much different score than when the proposed word is matched there independent of context. (Though provision has been made to allow context in word proposals, the appropriate mechanisms have not yet been enabled in the word matcher.)

A Controller governs the formation, evaluation and refinement of theories, essentially deciding who does what when, while keeping track of what has already been done and what is left to do. It can also take specific requests from one part of the system that another part be activated on

some specific job, but retains the option of when to act on each request. (In running SPEECHLIS with early versions of the control, syntactic and semantic components, we found several places where for efficiency, it was valuable for Syntax to be able to question Semantics directly during parsing. (N.B. We will be using initial capitals on the words "syntax", "semantics" and "pragmatics" when referring to parts of SPEECHLIS.) Thus, it is currently also possible for Syntax to make a limited number of requests directly to Semantics. How much more the initial control structure will be violated for efficiency's sake in the future is not now clear.)

The reason that processing does not stop after initial hypotheses have been formed about the utterance is that various _events_ may happen during the analysis of a theory which would tend to change SPEECHLIS's confidence in it, or to refine or modify it. For example, if no word could be matched just to the right of a given word match, we would be less certain about its being in the original utterance. On the other hand, in an utterance extracted from a discussion of the Apollo 11 moon rocks, if "sample" were to match well to the right of a word match for "lunar", we would be more confident about both words being in the original utterance. Entities called _Event Monitors_ are set up as active agents (i.e. demons) by the higher-level components to watch for events, and create appropriate _Notices_ when one has

occurred. Examples of semantic monitors and events will be found further on in this paper.

To summarize then, the semantics component of SPEECHLIS has available to it the following facilities from the rest of the system: access to the words which have been found to match some region of the acoustic input and information as to how close to the description of the input that match is; ability to ask for a word to be matched against some region of the input; and ability to build or flesh out theories based on its own knowledge and to study those parts of a theory built by Syntax and Pragmatics. Given this interface with the rest of the SPEECHLIS world, how does Semantics make its contribution to speech understanding, and what facets of the general understanding process does it allow us to study and hopefully contribute to solving?

B. How SPEECHLIS Semantics Works

The primary source of permanent semantic knowledge in SPEECHLIS is a network of nodes representing words, "multi-word names", concepts, specific facts, and types of syntactic structure. A network representation was chosen because the local and global semantic predictions about an utterance described earlier come from the associations among words and concepts in the domain and their possible surface realizations. Hanging onto each concept node is a frame

containing further information about its relations with the words and other concepts it is linked to, and which is also used in making predictions. The following sections describe how such predictions are enabled.

B.1 Network-based Predictions

Multi-Word Names

Each content word in the vocabulary (i.e. words other than articles, conjunctions, and prepositions. E.g. "ferric", "iron", "contain") is associated with a single node in the semantic network. From each word node, links go out to various other nodes. The first ones of interest in considering local predictions are those that go to nodes representing "multi-word names" of which the original word is a part. For example, "fayalitic olivine" is an multi-word name linked to both "fayalitic" and "olivine"; "fine-grained igneous rock" is one linked to the word "fine-grained" and the multi-word name "igneous rock".

Representing multi-word names in this way enables us to maintain a reasonable size dictionary in SPEECHLIS (i.e. by not having to make up compound entries like "fayalitic-olivine" and "principal-investigator") anc also to make local predictions. That is, any given word match may be partial evidence for a multi-word name of which it is a part. The remaining words may be in the word lattice,

adjacent and in the right order, or missing due to poor match quality. In the former case, one would eventually notice the adjacency and hypothesize (i.e. create a theory) that the entire multi-word name occurred in the original utterance. In the latter case, one would propose the missing words in the appropriate region of the word lattice, with a minimum acceptable match quality directly proportional to the urgency of the match's success. That, in turn, depends on how necessary it is for the word match to be part of a multi-word name. That is, given a word match for "oxide", Semantics would propose "ferrous" or "ferric" to its left (were neither in the word lattice), naming "ferrous oxide" or "ferric oxide". Given a match for "ferric" or "ferrous", Semantics would make a more urgent proposal for "oxide", were it not following in the word lattice, since neither word could appear in an utterance alone. Further details on the proposing and hypothesizing processes will be given below

There is another advantage to representing multi-word names in this way rather than as compound entries in the dictionary. As an immediate consequence, it turns out that fayalitic olivine is a type of olivine, a fine-grained igneous rock is a type of igneous rock which is a type of rock, and a principal investigator is a type of investigator. No additional links are needed to represent this class information for them.

Concept-Argument Relations

From the point of view of Semantics, an action or an event is a complex entity, tying several concepts together into one that represents the action or event itself. Syntactically, an action or event can be described in a single clause or noun phrase, each concept realizing some syntactic role in the clause or phrase One of these concepts is that associated with the verb or nominal (i.e. nominalized verb) which names the <u>relation</u> involved in the action or event. The other concepts serve as <u>arguments</u> to the relation. For a verb, this means they serve as its subject, object, etc.; for a nominal, it means they serve as pre-modifiers (e.g. adjectives, noun-noun modifiers, etc) or as post-modifiers (e.g. prepositional phrases, adverbials, etc.) For example,

```
John went to Santa Barbara in May.
SUBJ VERB    PREP PHRASE   PREP PHRASE

John's  trip to Santa Barbara in May.
PREMOD NOMINAL   PREP PHRASE  PREP PHRASE
```

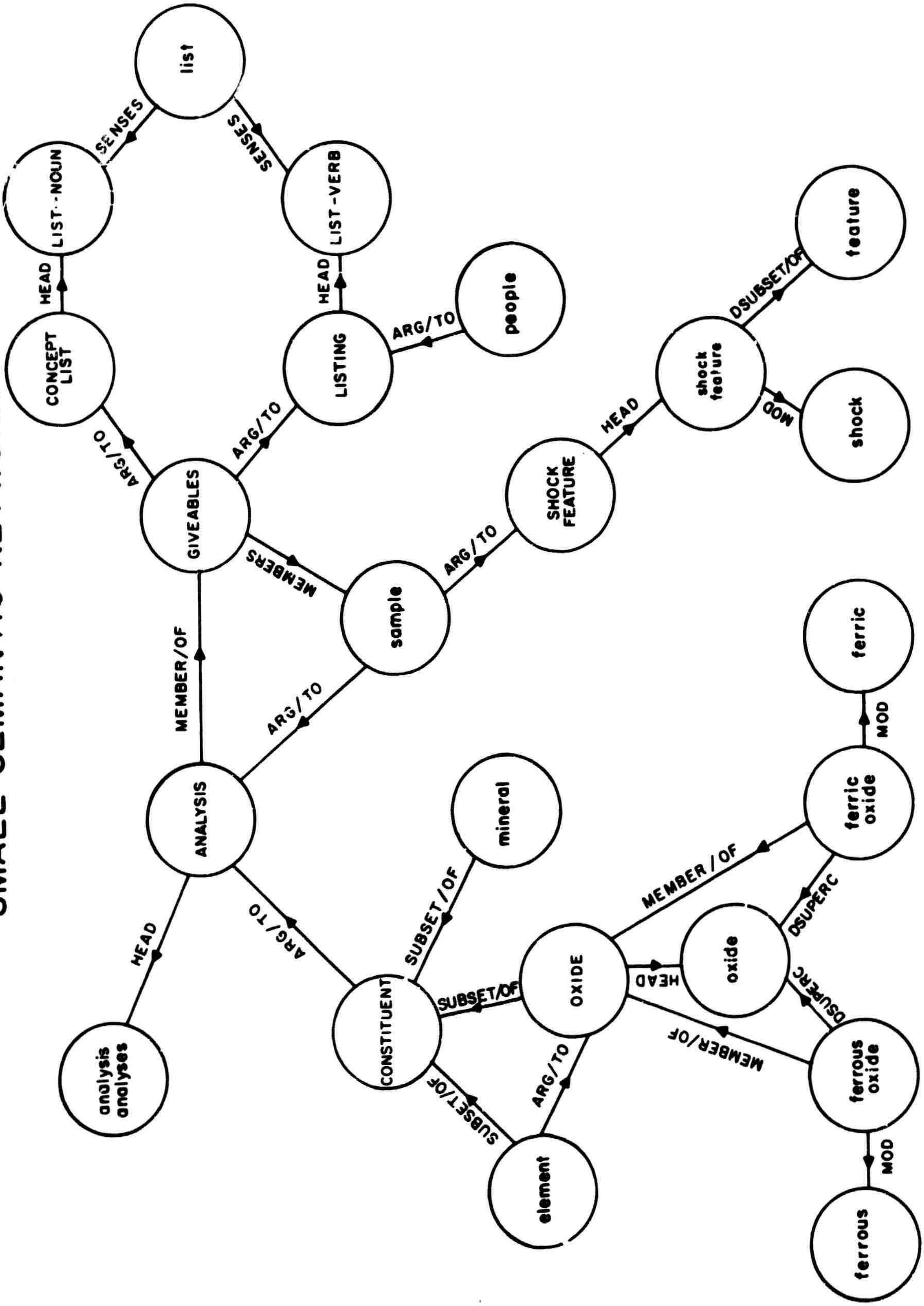
In the semantic network, an action or event concept is linked to the one which names the relation and the ones which can fill its arguments. This is another basis for network-based local predictions.

Semantics uses its knowledge of words, multi-word names, and concepts to make hypotheses (i.e. <u>theories</u>)

about possible local contexts for one or more word matches, detailing how the word matches fit into that context. Given a word match, Semantics follows links through the network, looking for multi-word names and concepts which it may partially instantiate. On each of the other components of the partially instantiated name or concept, Semantics sets monitors. Should a monitored node subsequently be instantiated (and conditions on the instantiation specified in the monitor be met), a notice is created, calling for the construction of a new, expanded theory.

To see this, consider the network shown in Figure 1 and a word match for "oxide". Since "oxide" occurs in the multi-word names "ferrous oxide" and "ferric oxide", Semantics would set monitors on the nodes for "ferrous" and "ferric", watching for either's instantiation to the immediate left of "oxide". It would also propose them there. Since the net shows that oxides can be constituents of rocks and a rock constituent can be one argument to the concept CONTAIN (the other argument being the concept SAMPLE), Semantics would also monitor the node for CONTAIN (and SAMPLE).

SMALL SEMANTIC NETWORK



Subsequently, an acceptable word match for "contain" or one of its inflected forms, or one which instantiates SAMPLE (e.g. "rock"), would be seen by a monitor and result in the creation of a notice linking "oxide" with the new word match.

Each notice has a weight representing how confident Semantics is that the resulting theory is a correct hypothesis about the original utterance. In the above, Semantics is less certain that a theory for "rock" and "oxide" will eventually instantiate the concept CONTAIN than will a theory for "contain" and "oxide". The event for the latter is given a higher weight than the former. (One is more certain that a particular relation has been expressed if one has heard its name mentioned rather than one or more of its possible arguments.)

Syntactic Structures

Nodes corresponding to the syntactic structures produced by the grammar (e.g. noun phrases, to-complements, relative clauses, etc.) are also used in making local predictions, several examples of which follow. First, if an argument to some concept can be specified as a particular syntactic structure with a particular set of syntactic features, we want to predict an occurrence of that structure, given an instantiation of the concept's head. For example, one of the things that the object of the

concept headed by "anticipate" may be is an embedded sentence whose tense is future to that of "anticipate" in the matrix sentence. We want to be able to predict and monitor for any such structures and notice them if built.

> I anticipate that we will make 5 trips to L.A.
> I anticipated that we would have made 5 trips to L.A. by November.

More generally, we want to be able to use any co-occurrence restrictions on lexical items and syntactic structures or features in making predictions. For example, when different time and frequency adverbials may be used depends on the mood, tense, and aspect of the main clause and certain features of the main verb. "Already", for instance, prefers that clauses in which it occurs, headed by a non-stative verb, be either perfective or progressive or both, unless a habitual sense is being expressed. E.g.

> John has already eaten 15 oysters.
> John is already sitting down.
> ?John already ate 15 oysters.
> (Perfective is preferable.)
> *John already sits down.
> John already runs 5 miles a day. (Habitual)

Secondly, if a concept with an animate agent as one of its arguments is partially instantiated, we want to predict an expression of the agent's purpose in the action. Now it is often possible to recognize "purpose" on syntactic grounds alone, as an infinitive clause introduced by "in order to", "in order for X to", "to" or "for X to". For

example,

> John's going to Stockholm to visit Fant's lab.
> I need $1000 to visit Tbilisi next summer.
> John will stay home in order for Rich to finish his paper.

These syntactic structure nodes then facilitate the search for a "purpose": they permit monitors to be set on the semantic concept of PURPOSE, which can look for, inter alia, the infinitive clauses popped by Syntax.

A third case of using syntactic structure nodes in making local predictions is almost as much a question of pragmatics as one of semantics. It seems to be more likely for a person to ask about a restricted set of objects than an entire set. Here we are talking about entire sets which are named by single English words like "rocks", "elements", "cities", etc., where restrictions are given syntactically in the form of pre-modifiers, prepositional phrases or relative clauses. For example,

> Tell me the cities.
>
> Tell me the cities which we have visited since September.

The first utterance is extremely unlikely.

B.2 Case Frame based Predictions

Description of a Case Frame

Additional information about how an action or event concept made up of a relation and its arguments may appear in an utterance is given in a case frame, a la Fillmore [4], associated with the concept. Case frames are useful both in making local predictions and in checking that some possible syntactic organization of the word matches in a theory supports Semantics' hypotheses. Figure 2 shows the case frames for the concepts ANALYSIS and CONTAIN.

A case frame is divided into two parts: the first part contains information relating to the case frame as a whole: the second, descriptive information about the cases. (In the literature, cases have been associated only with the arguments to a relation. We have extended the notion to include the relation itself as a case, specifically the head case (NP-HEAD or S-HEAD). This allows a place for the relation's instantiation in an utterance, as well as the instantiations of each of the arguments.)

CASE FRAME FOR ANALYSIS

```
(((REALIZES . NOUN-PHRASE))
  (NP-HEAD (EQU .14) NIL OBL)
  (NP-OBJ (MEM .1) (OF FOR) ELLIP)
  (NP-LOC (MEM .7) (IN FOR OF ON) ELLIP))
```

(a)

CASE FRAME FOR CONTAIN

```
(((REALIZES . CLAUSE)
  (ACTIVSUBJ S-LOC)
  (PASSIVSUBJ S-PAT))
  (S-HEAD (EQU.20) NIL OBL)
  (S-LOC (MEM .7) (IN) OBL)
  (S-PAT (MEM .1) NIL OBL))
```

(b)

CONCEPT 14 - CONCEPT OF ANALYSIS
CONCEPT 1 - CONCEPT OF COMPONENT
CONCEPT 7 - CONCEPT OF SAMPLE
CONCEPT 20 - CONCEPT OF CONTAIN

Figure 2

Among the types of information in the first part of the case frame is a specification of whether a surface realization of the case frame will be parsed as a clause or as a noun phrase, indicated in our notation as (REALIZES . CLAUSE) or (REALIZES . NOUN-PHRASE). If as a clause, further information specifies which cases are possible active clause subjects (ACTIVSUBJ's) and which are possible passive clause subjects (PASSIVSUBJ's). In the case of CONTAIN (Figure 2b), the only possible active subject is its location case (S-LOC), and the only possible passive subject is its patient case (S-PAT). For example,

```
Does each breccia contain olivine?
          S-LOC           S-PAT

Is olivine contained in each breccia?
    S-PAT                     S-LOC
```

(While not usual, there are verbs like "break" which allow several possible cases to become its active subject.

John broke the vase with a rock.
A rock broke the vase.
The vase broke.

However, which case actually does so falls out from which cases are present. In ACTIVSUBJ, the cases are ordered, so that the first one which occurs in an active sentence will be the subject. There is no syntactic preference, however, in selecting which case becomes passive subject, so the case names on PASSIVSUBJ are not ordered.) The first part of the case frame may also contain such information as inter-case restrictions, as would apply between instantiations of the

arguments to RATIO (i.e. that they be measurable in the same units).

The second part of the case frame contains descriptive information about each case in the frame.

a) its name, e.g. NP-OBJ, S-HEAD (The first part of the names gives redundant information about the frame's syntactic realization: "NP" for noun phrase and "S" for clause. The second part is an abbreviated Fillmore-type [4] case name: "OBJ" for object, "AGT" for agent, "LOC" for location, etc.)

b) the way it can be filled - whether by a word or phrase naming the concept (EQU) or by either's naming an instantiation of it (MEM), e.g. (EQU . SAMPLE) would permit "sample" or "lunar sample" to fill the case, but not "breccia". Breccia, by referring to a subset of the samples, only instantiates SAMPLE but does not name it.

c) a list of prepositions which could signal the case when it is realized as a prepositional phrase (PP). If the case were only realizable as a premodifier in a noun phrase or the subject or unmarked object of a clause, this entry would be NIL.

d) an indication of whether the case must be explicitly specified (OBL), whether it is optional and unnecessary (OPT), or whether, when absent, must be derivable from context (ELLIP). For example, in "The bullet hit.", the object case - what was hit - must be derivable from context in order for the sentence to be "felicitous" or well-posed. (We plan to replace this static, three-valued indication of sentence level binding with functions to compute the binding value. These functions will try to take into account such discourse level considerations as who is talking, how he talks and what aspects of the concept he is interested in.)

Uses of Case Frames

Semantics uses case frame information for making local predictions and checking the consistency of syntactic and

semantic hypotheses. These predictions mainly concern the occurrence of a preposition at some point in the utterance or a case realization's position in an utterance relative to cases already realized. The <u>strength</u> of such a prediction depends on its cost: the fewer the words or phrases which could realize the case, and the narrower the region of the utterance in which to look for one, the cheaper the cost of seeking a realization. Since there are fewer words and phrases which name a concept (EQU marker) as opposed to instantiating it (MEM marker), cases marked EQU would engender stronger predictions. (The process of localizing the prediction will be discussed further on.) The <u>urgency</u> of the prediction depends on its likelihood of success: if the case must be realized in the utterance (OBL marker), the prediction should be successful if the initial hypothesis about the concept associated with the case frame is correct. If the case need not be present in the utterance (ELLIP or OPT marker), even if the initial hypothesis is correct, the prediction need not be successful.

With respect to localizing case frame predictions, there are a number of simple strategies which, though not guaranteed successful, are rather helpful and inexpensive to employ. If a case can be realized as a premodifier of the head, we predict its realization to the immediate left of the head case. If it can be realized as a prepositional phrase, we predict one of its prepositions to the immediate

right of the head case and the realization of the case itself to the right of that. The obvious change in strategy is made for predicting the location of the head case.

Consider the case frame for ANALYSIS in Figure 2a for example. If we were to have a theory that the word "analysis" occurred in the utterance, we would predict (though not urgently because of the ELLIP markers) the following: 1) an instantiation of either COMPONENT or SAMPLE to its immediate left, 2) either "of" or "for" to its immediate right, followed by an instantiation of COMPONENT, and 3) either "in", "for", "of", or "on" to its immediate right, followed by an instantiation of SAMPLE. It doesn't matter that the above predictions are contradictory: if more than one prediction were successful (i.e. there were more than one way of reading that area of the speech signal), it would simply be the case that more than one refinement of the original theory for "analysis" would be made, each incorporating a different alternative. Further localization strategies include predicting possible subjects to the left of a hypothesized main verb (i.e. clause head) and possible objects to its right. If one has a hypothesis about the voice of the clause (i.e. active or passive), the number of predictions could be reduced.

It is important to remember here that in most cases we are predicting likely locations for case realizations, not

necessary ones. If they fail to appear in the places predicted, it does not cast doubts on a theory. English allows considerable phrase juggling -- e.g. preposing prepositional phrases, fronting questioned phrases, etc. And, of course, not all predicted pre- and post-modifiers of a noun can occur to its immediate left or right. This must be remembered in considering how these local, frame-based predictions can be employed. Leftness and rightness constraints are implemented in SPEECHLIS as additional requests associated with proposals and monitors. For example, consider a theory that the word "contain" occurs in an utterance. Under the hypothesis that the clause is active, we would include in the monitor set on the concept SAMPLE, the only possible active subject, that its instantiation be to the left of the match for "contain". In the monitor set on COMPONENT, the active object, we would indicate a preference for finding its instantiation to the right. This latter is only a preference because by question fronting, the object may turn up to the left. E.g. "What rare earth elements does each sample contain?". (Notice that regardless of where an instantiation of either SAMPLE or COMPONENT is found in the utterance, it will be noticed by the appropriate monitor. It is only the value of the particular concept instantiation to the theory setting the monitor that is affected by a positional preference.)

The process of checking the consistency of Syntax's and Semantics' hypotheses uses much the same information as that of making frame-based local predictions. As word matches are included in a theory, Semantics represents its hypotheses about their semantic structure in <u>case frame tokens</u>. These are instances of case frames which have been modified to show which word match or which other case frame token fills each instantiated case.

The two case frame tokens in Figure 3 represent semantic hypotheses about how the word matches for "analyses", "ferrous" and "oxide" fit together. "Analyses" is the head (NP-HEAD) of a case frame token whose object case (NP-OBJ) is filled by another case frame token representing "ferrous oxide". Another way of showing this is in the tree format of Figure 4.

CASE FRAME TOKENS

```
[Cft #6
    ((( Realizes . Noun-Phrase ))
     ( Np-Head (Analyses . 14 ) Nil Obl)
     ( Np-Goal (Cft#5 . 1) (Of For ) Ellip)
     ( Np-Loc (Mem . 7 ) ( In For Of On ) Ellip))]

[Cft #5
    ((( Realizes . Noun Phrase)
      ( Case of Cft #6 ))
     ( Np-Mod (Ferrous . 13) Nil Obl)
     ( Np-Head (Oxide . 5) Nil Obl )) ]
```

Figure 3

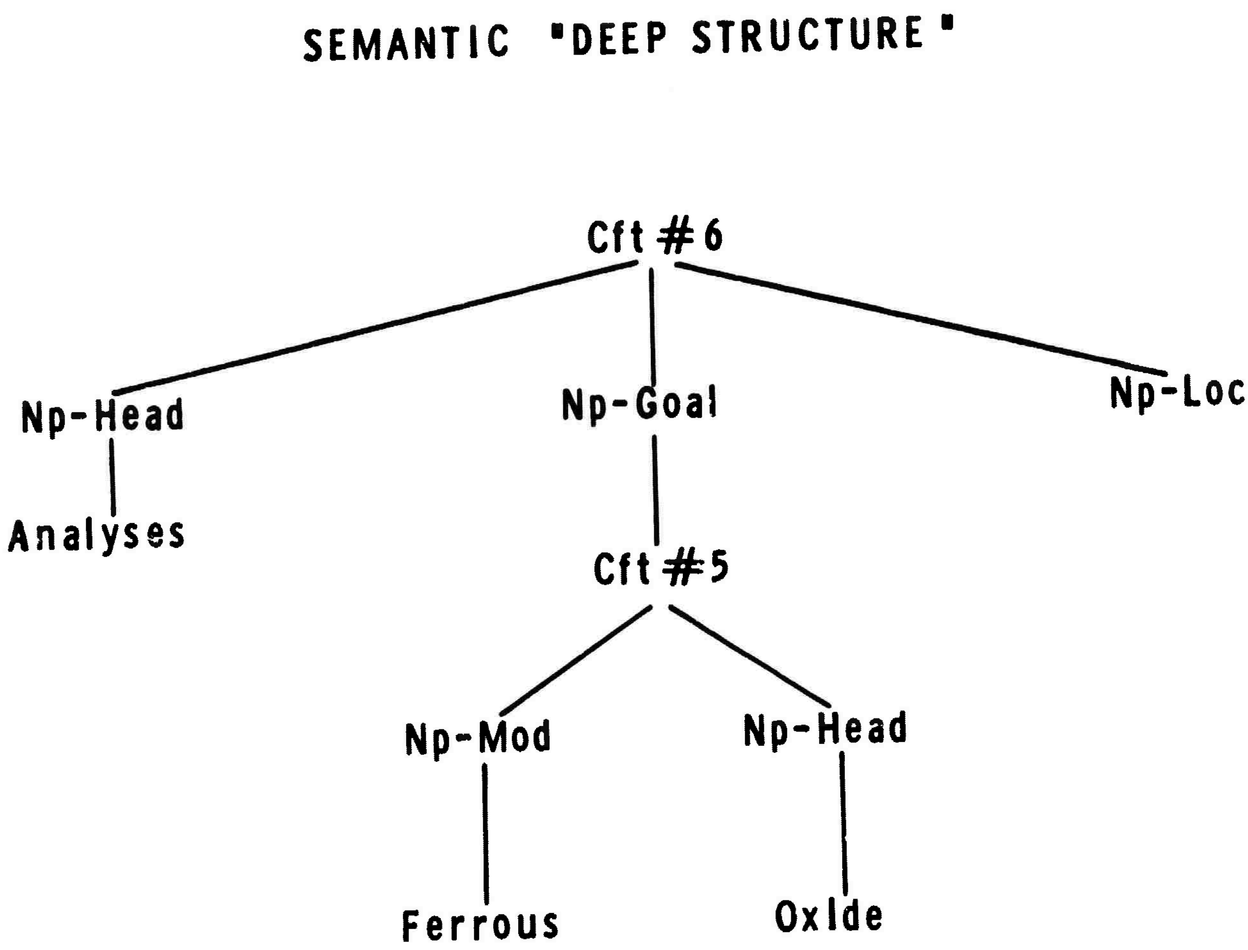

SEMANTIC "DEEP STRUCTURE"
Cft #6
Np-Head
Np-Goal
Np-Loc
Analyses
Cft #5
Np-Mod
Np-Head
Ferrous
Oxide

Figure 4

Semantics uses case frame tokens both to expedite Syntax's building structures consistent with its own hypotheses and to evaluate each of those it has built with respect to how many of its hypotheses have been substantiated and how much of each structure is inconsistent with, or irrelevant to, those hypotheses. We will consider the latter point first, using the case frame token in Figure 3a as an example. Syntactically, there are only a small number of ways of structuring this set of cases within the utterance: the head case must appear as the syntactic head and the object case must be realized as either in a prepositional phrase or relative clause or as an adjectival modifier on the head. Thus, in Figure 5, syntactic structures (a) and (b) would confirm the semantic hypotheses in Figure 3, while (c), where "analyses" modifies "oxide", would not and would therefore receive a lower evaluation. Notice that the only difference between the terminal strings of (a) and (c) is the presence of the preposition "of". It only takes a small, acoustically ambiguous, word to make the difference between an acceptable syntactic structure and an unacceptable one. Yet, given the two surrounding words and asked to test for the presence of a specific such function word, the acoustics component should be able to do it.

SYNTACTIC STRUCTURES

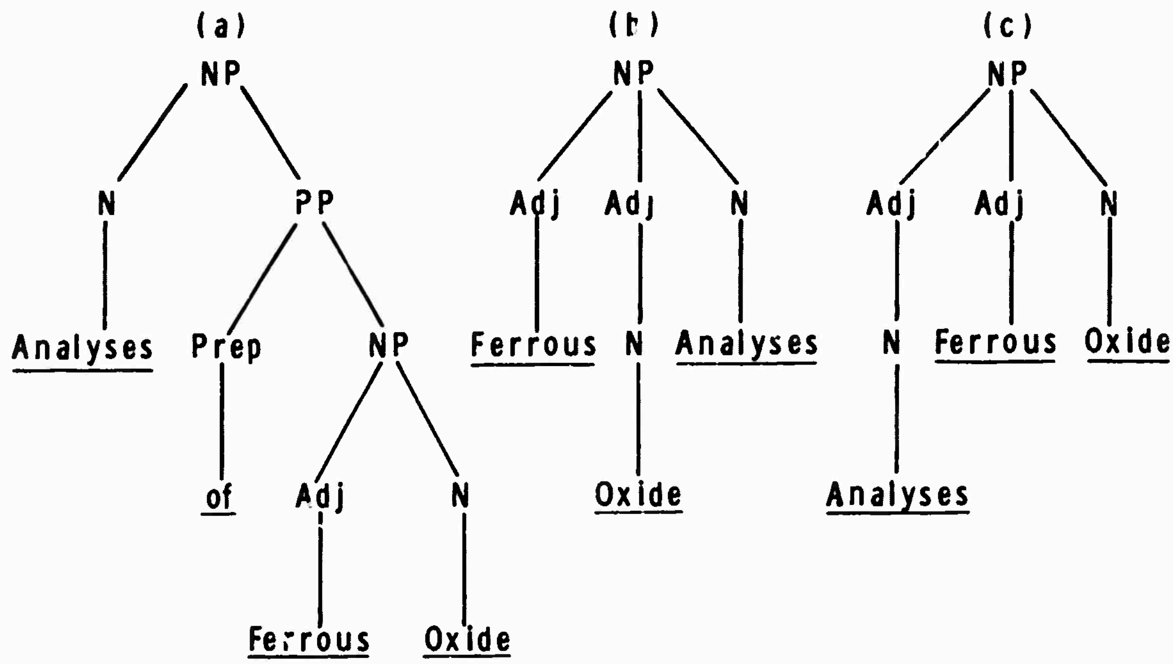


Figure 5

With respect to expediting Syntax's building a structure consistent with Semantics's hypotheses, the point is that Syntax should not make random choices in places

where Semantics has information that can be used to order them. This is implemented via Syntax's ability to ask questions of Semantics on the arcs of the Transition Network Grammar [2, 21]. For example, noun/present-participle/noun strings may have the structure of a preposed relative clause like "the olivine containing sample" (i.e. "the sample which contains olivine") or a reduced relative clause like "the sample containing olivine" (It may be that prosodies help distinguish these two types of relative clauses in spoken utterances, but, as we suggested earlier, it may also be the case that this additional cue is not used if the phrase is already disambiguated by semantics or context.)

In parsing the string "the olivine containing sample", syntax must chose whether it is indicative of a preposed relative clause or a reduced one. If preposed, "olivine containing" would have the structure shown in Figure 6a, with "olivine" as object and subject unknown. This is acceptable to Semantics, since olivine, a mineral, may be both container and containee. "Sample" then becomes the head of the noun phrase and simultaneously the subject of the preposed relative clause, as shown in Figure 6b. This semantics accepts. Were the word match for "sulfur" instead of "sample", the final structure -- "the sulfur which contains olivine" -- would be semantically anomalous, and Semantics would advise the parser not to pursue this path. On the other hand, "sample containing", with "sample" as

object (Figure 6c), is semantically anomalous in the lunar rocks domain, so again the parser would be advised not to pursue this path further.

The opposite happens when the parser considers both strings as normal relative clauses. "The olivine containing sample" has the intermediate structure shown in Figure 7a, which is as bad as in 6c above. Only "The sample containing olivine" is reasonable as a normal reduced relative clause (Figures 7b and 7c).

The olivine containing sample The sample containing olivine

Figure 6

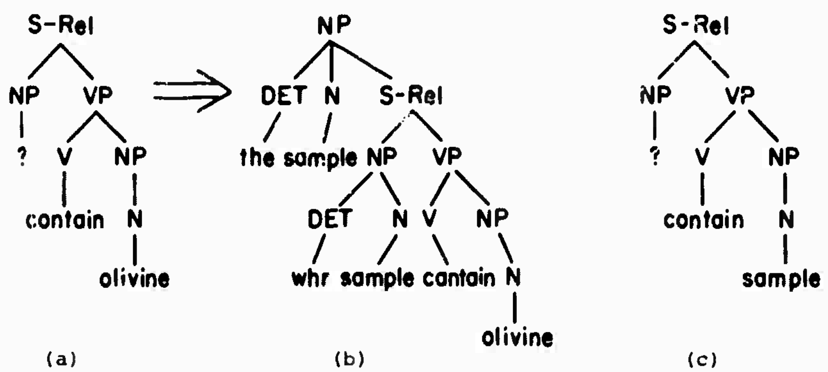
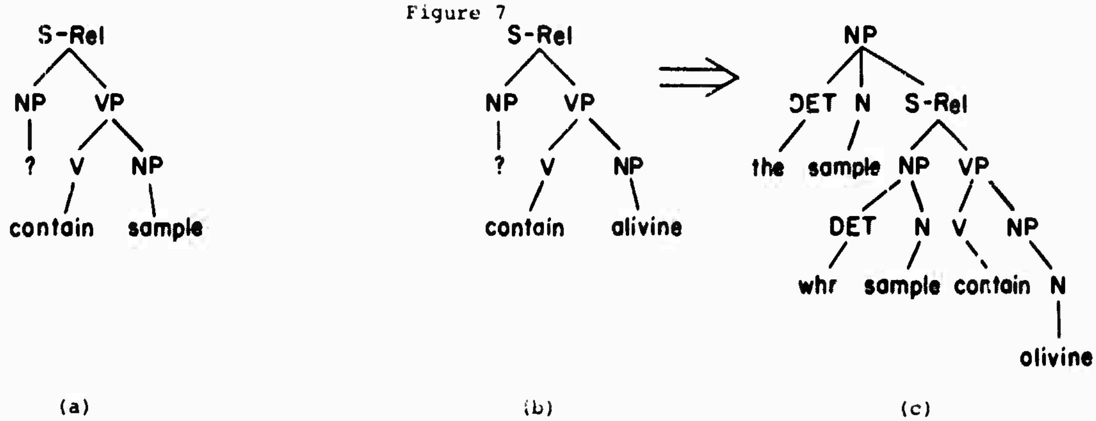


Figure 7



B.3 Further Tasks

Shortly before this paper was finished, we had begun to look at how to represent the semantic knowledge that SPEECHLIS should have with respect to cross-utterance phenomena (i.e. making global predictions and resolving anaphora and ellipsis), but we had not come to any definite conclusions. Another task of Semantics in the process of being implemented was that of turning the hypothesized utterance into a formal procedure for operating on its data base in order to answer questions or to absorb new information. One half of the problem, the interpretation process, has been done once in LUNAR [24] and should follow similarly here. The other half, structuring the knowledge for assimilation of new facts, is difficult and other papers in this volume have more to say about this issue.

To conclude this section, Semantics is used in SPEECHLIS in several ways to aid the general speech understanding task. 1) It makes predictions local to a single utterance. 2) It collects sets of word matches which substantiate its hypotheses about the meaning of the utterance. 3) It checks the possible syntactic organizations of the word matches for confirmation or discrediting of those hypotheses. This it does using both a semantic network representing, inter alia, the concepts known in the domain and the words and multi-word names

available for expressing them, and also *case frames* which give further information about their surface and syntactic realization.

VI. SUMMARY AND CONCLUSIONS

We attempted in this paper to make two major points. First, a listener requires a knowledge of meaningful concepts and their possible surface realizations in order to recover a speaker's intended utterance. This knowledge is applied to making both local and global predictions about the current utterance and future ones and also to checking the meaningfulness of hypothesized reconstructions of that utterance. This constrains the many possible ways of hearing any given speech signal. Secondly, we made the point that speech is a good context to study understanding, mainly because it forces us to confront and deal with aspects of understanding that either do not arise or could be circumvented in understanding written text. We strongly believe in the value of studying automatic speech understanding, as it cannot help but give us further insight into language and possibly even our own language use.

REFERENCES

[1] Barnett, J., "A Vocal Data Management System", IEEE Transactions on Audio and Electroacoustics, Vol. Au-21, No.3, June 1973, pp. 185-188.

[2] Bates, M., "The Use of Syntax in a Speech Understanding System," Proc. IEEE Symp. Speech Recognition, CMU, April 1974.

[3] Bolinger, D., "Accent is Predictable (if you're a Mind Reader)", Language 48(3), 1972, pp. 633-644.

[4] Fillmore, C., "The Case for Case" in Bach and Harms, Universals in Linguistic Theory, pp. 1-90 (1968).

[5] Forgie, J.W., "Speech Understanding Systems", Semiannual Technical Summary, 1 December 1973 - 31 May 1974. Lincoln Laboratory, MIT.

[6] Fromkin, V., "The Non-anomalous Nature of Anomalous Utterances", Language, 47(1), pp. 27-53, March 1971.

[7] Marcus, M., "Wait-and-See Strategies for Parsing Natural Language", MIT Artificial Intelligence Laboratory Working Paper 75, August 1974.

[8] Neely, R.B., "On the Use of Syntax and Semantics in a Speech Understanding System", Department of Computer Science, Carnegie-Mellon University, May 1973.

[9] Newell, A. et.al, Speech Understanding Systems: Final Report of a Study Group, North-Holland Publishing Company, Amsterdam, Netherlands, 1973.

[10] Olson, H.F. and H. Belar, "Phonetic Typewriter", Journal of the Acoustic Society of America, 28(6), November 1956, pp. 1072-1081.

[11] Pollack, I. and J. Pickett, "The Intelligibility of Excerpts from Conversation", Language and Speech, VI, pp. 165-171 (1964).

[12] Reddy, D.R., L.D. Erman, R.D. Fennell, and R.B. Neely, "The HEARSAY Speech Understanding System: An Example of the Recognition Process", Proceedings of the 3rd International Joint Conference on Artificial Intelligence, Stanford University, August 1973, pp. 185-193.

[13] Riesbeck,C.K., "Computational Understanding: Analysis of sentences and context", Ph.D. Thesis, Stanford University, 1974. [Reprinted in part in Schank, R. (ed), Conceptual Information Processing, ------ 1974.]

[14] Ritea, H.B., "A Voice-controlled Data Management System", Proc. IEEE Symp. Speech Recognition, CMU, April 1973.

[15] Rovner, P., B. Nash-Webber and W.A. Woods, "Control Concepts in a Speech Understanding System", BBN Report No. 2703, Bolt Beranek and Newman Inc., Cambridge, Ma. (1973). (Also in Proc. IEEE Symp. Speech Recognition, CMU, April 1973.)

[16] Sussman, G. , "Some Aspects of Medical Diagnosis", MIT Artificial Intelligence Laboratory, Working Paper 56, December 1973.

[17] Walker, D.E., "Speech Understanding, Computational Linguistics and Artificial Intelligence", SRI Artificial Intelligence Center, Technical Note 85, August 1973.

[18] Walker, D.E., "Speech Understanding through Syntactic and Semantic Analysis", Proceedings of the 3rd International Joint Conference on Artificial Intelligence, Stanford University, August 1973, pp. 208-215.

[19] Wanner, E., "Do We Understand Sentences from the Outside-In or from the Inside-Out", Daedelus, Summer 1973, pp. 163-183.

[20] Winograd, T., "PROGRAMMAR: A language for writing grammars", MIT Artificial Intelligence Laboratory Memo No. 181, November 1969.

[21] Winograd, T., "Procedures as Representation for Data in a Computer Program for Understanding Natural Language", Report MAC-TR-84, MIT Project MAC, Cambridge, Massachusetts (February 1971). Published as Understanding Natural Language (Academic Press, New York, New York, 1972).

[22] Woods, W.A., "Transition Network Grammars for Natural Language Analysis", Communications of the ACM, 13(10), October 1970. pp. 591-606.

[23] Woods, W.A., "Motivation and Overview of BBN SPEECHLIS: An Experimental Prototype for Speech Understanding Research," Proc. IEEE Symp. Speech Recognition, CMU, April 1974.

[24] Woods, W.A., R.M. Kaplan, and B. Nash-Webber, "The Lunar Sciences Natural Language Information System: Final Report", BBN Report No. 2388, Bolt Beranek and Newman Inc., Cambridge, Ma. (June 1982).

[25] Woods, W.A. and J.I. Makhoul, "Mechanical Inference Problems in Automatic Speech Understanding", Artificial Intelligence, Vol. 5, No. 1, pp. 73-92 (Spring 1974).